Multidimensional two-component Gaussian mixtures detection

Béatrice Laurent, Clément Marteau, Cathy Maugis-Rabusseau

Multidimensional two-component Gaussian mixtures detection

Béatrice Laurent and Clément Marteau and Cathy Maugis-Rabusseau

Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse
INSA de Toulouse,
135, avenue de Rangueil,
31077 Toulouse Cedex 4, France.

Abstract: Let $(X_1, \ldots, X_n)$ be a d-dimensional i.i.d sample from a distribution with density f. The problem of detection of a two-component mixture is considered. Our aim is to decide whether f is the density of a standard Gaussian random d-vector ($f = \phi_d$) against f is a two-component mixture: $f = (1-\varepsilon)\phi_d + \varepsilon\phi_d(. - \mu)$ where (ε, μ) are unknown parameters. Optimal separation conditions on ε, μ, n and the dimension d are established, allowing to separate both hypotheses with prescribed errors. Several testing procedures are proposed and two alternative subsets are considered.

AMS 2000 subject classifications: Primary 62H15; secondary 62G30.
Keywords and phrases: Gaussian mixtures, Non-asymptotic testing procedure, Order statistics, Separation rates.

1. Introduction

Let $\underline{X} = (X_1, \ldots, X_n)$ be an i.i.d n-sample, where for all $i \in \{1, \ldots, n\}$, X_i corresponds to a d-dimensional random vector, whose distribution admits a density f w.r.t the Lebesgue measure on $\mathbb{R}^d$. In the following, we denote by $\phi_d(.)$ the density function of the standard Gaussian distribution $\mathcal{N}_d(0_d, I_d)$ on $\mathbb{R}^d$. Our aim is to test

$$H_0 : f = \phi_d \quad \text{against} \quad H_1 : f \in \mathcal{F}, \tag{1}$$

where

$$\mathcal{F} = \left\{ f_{(\varepsilon,\mu)} : x \in \mathbb{R}^d \mapsto (1-\varepsilon)\phi_d(x) + \varepsilon\phi_d(x-\mu); \varepsilon \in]0,1[, \mu \in \mathbb{R}^d \right\}$$

is the set of two-component Gaussian mixtures on $\mathbb{R}^d$. Mixture models are at the core of several studies and provide a powerful paradigm that allows to model several practical phenomena. We refer to McLachlan and Peel (2000) for an extended introduction to this topic.

The particular case of a two-component mixture is sometimes referred as a contamination model. In some sense, a proportion ε of the sample is driven from a (Gaussian) distribution centered in μ while the remaining part of the data is centered. In this context, the testing problem (1) amounts to the detection of a plausible contamination inside the data at hand w.r.t. the null distribution. We refer for instance to Donoho and Jin (2004) for practical motivations regarding this problem. We stress that Gaussian mixture is at the core of our contribution since it provides a benchmark model for several practical applications. However, the results proposed in this paper could be certainly extended to a wide range of alternative distributions.

In a unidimensional setting ($d = 1$), the testing problem (1) has been widely considered in the literature in the last two decades. A large attention has been payed to methods based on the likelihood ratio, see e.g. Chernoff and Lander (1995), Azaïs et al. (2009) or Garel (2007). Concerning the construction of optimal separation conditions on the parameters (ε, μ), we can mention the seminal contribution of Ingster (1999). These conditions have been reached by the higher-criticism procedure proposed in Donoho and Jin (2004) in a specific *sparse* context, i.e. when $\varepsilon \ll 1/\sqrt{n}$ as $n \to +\infty$. Then, several extensions of this contribution have been proposed in an extended context: we mention for instance Cai et al. (2007) for a study including confidence sets and the *dense* setting ($\varepsilon \gg 1/\sqrt{n}$ as $n \to +\infty$), Cai et al. (2011) for heterogeneous and heteroscedastic mixtures, or Cai and Wu (2014) where general distributions and separation conditions have been investigated. In a slightly different spirit, a procedure based on the order statistics and non-asymptotic investigations on the testing problem (1) have been proposed in Laurent et al. (2014).

In the contributions mentioned above, only unidimensional distributions are considered. In a different setting (signal detection), multidimensional problems have been at the core of recent investigations. We mention e.g. Arias-Castro et al. (2011) or Butucea and Ingster (2013) among others. In a recent paper, Verzelen and Arias-Castro (2014) address the problem of testing normality in a multidimensional framework. They consider two-component Gaussian mixture alternatives where the proportions are fixed and the difference in means are sparse. However, up to our knowledge, the multidimensional testing problem as displayed in (1) has never been studied so far. We stress that in our setting, the proportion ε is allowed to depend on the number of observations n. The present paper proposes a first attempt in this context. Our aim is two-fold: we establish *optimal* separation conditions on the parameters (ε, μ) for the testing problem (1) in a first time, and describe the influence of the dimension d on the corresponding problem. In the same time, we propose various testing procedures and compare their theoretical performances.

In this paper, we assume that the norm of the mean parameter μ is bounded on

the alternative H_1. Given $M \in \mathbb{R}_+^*$, we deal with the subsets $\mathcal{F}_2[M]$ and $\mathcal{F}_\infty[M]$ defined as

$$\mathcal{F}_2[M] = \left\{ f_{(\varepsilon,\mu)} \in \mathcal{F};\ \varepsilon \in]0,1[, \|\mu\| \leq M \right\},$$

and

$$\mathcal{F}_\infty[M] = \left\{ f_{(\varepsilon,\mu)} \in \mathcal{F};\ \varepsilon \in]0,1[, \|\mu\|_\infty \leq M \right\},$$

where for a given $\mu \in \mathbb{R}^d$, $\|\mu\| = \left(\sum_{j=1}^d \mu_j^2\right)^{1/2}$ denotes the l_2-norm and $\|\mu\|_\infty = \max_{j=1\ldots d} |\mu_j|$ corresponds to the l_∞-norm. In this context, our main results can be gathered in the following theorem.

Theorem 1. *Let $\alpha, \beta \in]0,1[$ be fixed and $\mathcal{C}_+ = \mathcal{C}_+(\alpha,\beta,M)$, $\mathcal{C}_- = \mathcal{C}_-(\alpha,\beta,M)$ be two explicit constants. Then, there exists a level-α testing procedure $\tilde{\Psi}_{\alpha,2}$ such that*

$$\sup_{\substack{f \in \mathcal{F}_2[M] \\ \varepsilon\|\mu\| > \mathcal{C}_+ d^{1/4}/\sqrt{n}}} \mathbb{P}_f(\tilde{\Psi}_{\alpha,2} = 0) \leq \beta \quad \text{and} \quad \inf_{\Psi_\alpha} \sup_{\substack{f \in \mathcal{F}_2[M] \\ \varepsilon\|\mu\| > \mathcal{C}_- d^{1/4}/\sqrt{n}}} \mathbb{P}_f(\Psi_\alpha = 0) > \beta,$$

where the infimum is taken on all possible level-α testing procedures Ψ_α.

Similarly, there exists a level-α testing procedure $\tilde{\Psi}_{\alpha,\infty}$ such that

$$\sup_{\substack{f \in \mathcal{F}_\infty[M] \\ \varepsilon\|\mu\|_\infty > \mathcal{C}_+\sqrt{\ln(d)/n}}} \mathbb{P}_f(\tilde{\Psi}_{\alpha,\infty} = 0) \leq \beta \quad \text{and} \quad \inf_{\Psi_\alpha} \sup_{\substack{f \in \mathcal{F}_\infty[M] \\ \varepsilon\|\mu\|_\infty > \mathcal{C}_-\sqrt{\ln(d)/n}}} \mathbb{P}_f(\Psi_\alpha = 0) > \beta.$$

Theorem 1 indicates that the detection boundary associated to the testing problem (1) for the alternative subset $\mathcal{F}_2[M]$ is of order $d^{1/4}/\sqrt{n}$: detection is impossible (with a prescribed level β) if $\varepsilon\|\mu\|$ is smaller than $d^{1/4}/\sqrt{n}$, up to some constant. In the case of the alternative set $\mathcal{F}_\infty[M]$, the detection boundary depends on $\sqrt{\ln(d)}$. In these two cases, we propose optimal testing strategies in Sections 3.1, 3.2 and 3.3.

The paper is organized as follows. Two different lower bounds are proposed in Section 2 for both subsets $\mathcal{F}_2[M]$ and $\mathcal{F}_\infty[M]$ and proved in Section 5. Associated upper bounds are established in Section 3 and proved in Section 6. To this end, we will investigate the performances of three different testing procedures. Some useful lemmas are gathered in Appendix A, while Appendix B contains some technical results.

All along the paper, we use the following notations. For any density g on $\mathbb{R}^d$, we denote respectively by $\mathbb{P}_g$ and $\mathbb{E}_g$ the probability and expectation under the assumption that the common density of each X_i in the i.i.d. sample $(X_1, \ldots, X_n)$ is g. In the particular case where the $X_1, \ldots, X_n$ are i.i.d. with common density ϕ_d, which is associated to the null hypothesis H_0, we write $\mathbb{P}_0 := \mathbb{P}_{\phi_d}$ and $\mathbb{E}_0 := \mathbb{E}_{\phi_d}$. A

testing procedure Ψ denotes a measurable function of the sample $\underline{X}$, having values in $\{0,1\}$. By convention, we reject (resp. do not reject) H_0 is $\Psi = 1$ (resp. $\Psi = 0$). Given $\alpha \in]0,1[$, the test Ψ is said to be of level α if $\mathbb{P}_0(\Psi = 1) \leq \alpha$. In such a case, we write $\Psi = \Psi_\alpha$.

2. Lower bounds

2.1. Lower bound for the alternative class $\mathcal{F}_2[M]$

The non asymptotic minimax separation rates have been introduced by Baraud (2002). Let us recall the main definitions. Given $\beta \in]0,1[$, the class of alternatives $\mathcal{F}_2[M]$ and a level-α test Ψ_α, we define the uniform separation $\rho(\Psi_\alpha, \mathcal{F}_2[M], \beta)$ of Ψ_α over the class $\mathcal{F}_2[M]$ as the smallest positive number ρ such that the test has a second kind error at most equal to β for all alternatives $f_{(\varepsilon,\mu)}$ in $\mathcal{F}_2[M]$ such that $\varepsilon\|\mu\| \geq \rho$. More precisely,

$$\rho(\Psi_\alpha, \mathcal{F}_2[M], \beta) = \inf\left\{ \rho > 0;\ \sup_{\substack{f \in \mathcal{F}_2[M] \\ \varepsilon\|\mu\| \geq \rho}} \mathbb{P}_f(\Psi_\alpha = 0) \leq \beta \right\}.$$

Then, the (α,β)-minimax separation rate over $\mathcal{F}_2[M]$ is defined as

$$\underline{\rho}(\mathcal{F}_2[M], \alpha, \beta) = \inf_{\Psi_\alpha} \rho(\Psi_\alpha, \mathcal{F}_2[M], \beta),$$

where the infimum is taken over all level-α tests Ψ_α.

Theorem 2 proposes a lower bound for the minimax separation rate $\underline{\rho}(\mathcal{F}_2[M], \alpha, \beta)$. The main ingredient for the proof (displayed in Section 5) is the construction of particular distributions for which the separation of both hypotheses H_0 and H_1 will be impossible with a prescribed level β.

Theorem 2. *Let $\alpha, \beta \in]0,1[$ such that $\alpha + \beta < 0.29$. Define*

$$\rho^\# = \frac{1}{2\sqrt{C(M)}} \frac{d^{1/4}}{\sqrt{n}}, \quad \text{where} \quad C(M) = 1 + \frac{M^2}{2} e^{M^2}.$$

Then, if $\rho < \rho^\#$,

$$\inf_{\Psi_\alpha} \sup_{\substack{f \in \mathcal{F}_2[M] \\ \varepsilon\|\mu\| \geq \rho}} \mathbb{P}_f(\Psi_\alpha = 0) > \beta, \tag{2}$$

where the infimum is taken over all level-α tests. In particular, this implies that

$$\underline{\rho}(\mathcal{F}_2[M], \alpha, \beta) \geq \rho^\#.$$

Equation (2) indicates that the hypotheses H_0 and H_1 cannot be separated with prescribed first and second kind errors α and β following the value of the terms $\varepsilon, \|\mu\|, d$ and n. In particular, for any level-α testing procedure, one can find a distribution $f \in \mathcal{F}_2[M]$ such that $\varepsilon\|\mu\| \geq \rho^{\#}$ and $\mathbb{P}_f(\Psi_\alpha = 0) > \beta$. This result is obtained thanks to an assumption on the levels α and β. This assumption is essentially technical and could be removed thanks to additional technical algebra.

The condition $\varepsilon\|\mu\| \gtrsim d^{1/4}/\sqrt{n}$ is quite informative. First of all, since $\|\mu\|$ is bounded, the proportion parameter ε should be at least of order $1/\sqrt{n}$. This condition is often characterized as the *dense* regime in the literature. In the same time, the 'energy' $\|\mu\|$ should not be to small if one expects to detect a potential contamination in the sample. It is worth pointing out that Theorem 2 precisely quantifies the role played by the dimension d of the problem at hand. We will see in Section 3 that this lower bound is optimal, up to some constant.

2.2. Lower bound for the alternative class $\mathcal{F}_\infty[M]$

In this section, we concentrate our attention on the alternative $\mathcal{F} = \mathcal{F}_\infty[M]$. As in Section 2.1, we consider the (α, β)-minimax separation rate over $\mathcal{F}_\infty[M]$ defined as

$$\underline{\rho}(\mathcal{F}_\infty[M], \alpha, \beta) = \inf_{\Psi_\alpha} \rho(\Psi_\alpha, \mathcal{F}_\infty[M], \beta),$$

where the infimum is taken over all level-α tests Ψ_α and

$$\rho(\Psi_\alpha, \mathcal{F}_\infty[M], \beta) = \inf \left\{ \rho > 0; \sup_{\substack{f \in \mathcal{F}_\infty[M] \\ \varepsilon\|\mu\|_\infty \geq \rho}} \mathbb{P}_f(\Psi_\alpha = 0) \leq \beta \right\}.$$

Theorem 3 provides a lower bound for the minimax separation rate in this context. The proof is postponed to Section 5.

Theorem 3. *Let $\alpha, \beta \in]0, 1[$ such that $\alpha + \beta < 1$. Let*

$$\rho^\star = \sqrt{\frac{1}{C(M)} \frac{1}{n} \ln\left[1 + d\eta(\alpha, \beta)^2\right]}$$

where $C(M) = (1 + M^2 e^{M^2}/2)$ and $\eta(\alpha, \beta) = 2(1 - \alpha - \beta)$. Then, if $\rho < \rho^\star$,

$$\inf_{\Psi_\alpha} \sup_{\substack{f \in \mathcal{F}_\infty[M] \\ \varepsilon\|\mu\|_\infty \geq \rho}} \mathbb{P}_f(\Psi_\alpha = 0) > \beta.$$

This implies that

$$\underline{\rho}(\mathcal{F}_\infty[M], \alpha, \beta) \geq \rho^\star.$$

Theorem 3 indicates that the detection condition on the parameters (ε, μ) is affected by the change of the reference norm. In particular, the dependency w.r.t. the dimension of the data is in this context of order $\sqrt{\ln(d)}$ which makes in some sense the detection problem easier than the one considered in Section 2.1.

3. Upper bounds

In Section 2, we have proposed lower bounds on the separation region for the testing problem (1). In particular, we have proved that in some specific cases, related to the value of the parameters (ε, μ), testing is impossible, i.e. every level-α tests will be associated to a second kind error greater than a prescribed level β.

The aim of this section is to complete this discussion with upper bounds on the separation region. We propose three different testing procedures and investigate their related performances. In particular, we prove that these procedures reach the lower bounds presented above for both alternatives $\mathcal{F}_2[M]$ and $\mathcal{F}_\infty[M]$.

3.1. First testing procedure for the alternative class $\mathcal{F}_2[M]$

In a first time, we propose a procedure in the case where the alternative is expressed through the l_2-norm of μ. This procedure is based on the fluctuations of the empirical mean of the data. Intuitively, $\mathbb{E}_f[X] = \varepsilon\mu$ for all random vectors having density $f \in \mathcal{F}_2[M]$, while $\mathbb{E}[X] = 0$ under H_0. In particular, if the empirical mean of the sample has a large norm, there is a chance that the data have been driven w.r.t. a density f that belongs to $\mathcal{F}_2[M]$.

More precisely, given $\underline{X} = (X_1, \ldots, X_n)$, set $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let υ_α denote the $(1-\alpha)$ quantile of a chi-squared distribution with d degrees of freedom and define the test $\Psi_{1,\alpha}$ as

$$\Psi_{1,\alpha} = \mathbb{1}_{\{\|\sqrt{n}\bar{X}_n\|^2 > \upsilon_\alpha\}}. \tag{3}$$

The following theorem investigates the performances of this test.

Theorem 4. *Let $\alpha, \beta \in]0,1[$ be fixed. Then, the testing procedure $\Psi_{1,\alpha}$ introduced in (3) is of level α. Moreover, there exists a positive constant $C(\alpha, \beta, M)$ depending only on α, β and M such that*

$$\sup_{\substack{f \in \mathcal{F}_2[M] \\ \varepsilon\|\mu\| \geq \rho}} \mathbb{P}_f\left(\Psi_{1,\alpha} = 0\right) \leq \beta,$$

for all $\rho \in \mathbb{R}_+^*$ *such that*

$$\rho \geq C(\alpha, \beta, M)\frac{d^{1/4}}{\sqrt{n}}.$$

The above result indicates that the test $\Psi_{1,\alpha}$ is powerful as soon as $f \in \mathcal{F}_2[M]$ with $\varepsilon\|\mu\| \gtrsim d^{1/4}/\sqrt{n}$. According to the lower bound displayed in Theorem 2, it appears that the minimax detection frontier is of order $d^{1/4}/\sqrt{n}$ up to a constant, i.e. there exist $\mathcal{C}_-$ and $\mathcal{C}_+$ such that

- the hypotheses H_0 and H_1 cannot be separated if $\varepsilon\|\mu\| \leq \mathcal{C}_- d^{1/4}/\sqrt{n}$,
- there exists a level-α powerful test as soon as $\varepsilon\|\mu\| \geq \mathcal{C}_+ d^{1/4}/\sqrt{n}$.

These two assertions together provide the first part of Theorem 1. We stress that we do not investigate the value of the optimal constant associated to this separation problem ($\mathcal{C}_-$ and $\mathcal{C}_+$ do not match). Such a study indeed requires advanced asymptotic tools (see e.g. Ingster and Suslina, 2003) and is outside the scope of the paper.

3.2. Second testing procedure for the alternative class $\mathcal{F}_2[M]$

In a unidimensional context, Laurent et al. (2014) have introduced a testing procedure based on the ordered statistics. Such variables can indeed provide valuable informations on the deviation of the sample w.r.t. a benchmark distribution (e.g. under H_0). Although the ordered statistics of the sample are not clearly defined in a multidimensional setting, we can still project the data on a given axis and apply the procedure detailed in Laurent et al. (2014).

To this end, we split the sample $\underline{X}$ in two different parts $\underline{A} = (A_1, \ldots, A_{n/2})$ and $\underline{Y} = (Y_1, \ldots, Y_{n/2})$. For the sake of convenience, we assume without loss of generality that n is even and write $n/2 = n$ in the sequel. Set

$$v_n = \frac{\bar{A}_n}{\|\bar{A}_n\|} \text{ where } \bar{A}_n = \frac{1}{n}\sum_{i=1}^{n} A_i. \tag{4}$$

The axis generated by the vector v_n indicates the direction of the empirical mean of the sample $\underline{A}$. Under H_1, it provides an information on the direction where the contamination in the sample has occurred. Then, we project the remaining data $\underline{Y}$ on this axis. To this end, define

$$Z_i^{(v_n)} = \langle Y_i, v_n \rangle \quad \forall i \in \{1, \ldots, n\}.$$

The corresponding sample $\underline{Z}^{(v_n)} = (Z_1^{(v_n)}, \ldots, Z_n^{(v_n)})$ is unidimensional and we remark that under H_0, $Y_i \sim \mathcal{N}_d(0_d, I_d)$, hence conditionally on v_n, $Z_1^{(v_n)}, \ldots, Z_n^{(v_n)}$ are i.i.d. standard Gaussian variables since $\|v_n\| = 1$. Therefore, we can apply the test introduced in Laurent et al. (2014). Recall that this test is based on the ordered statistics $Z_{(1)} \leq \cdots \leq Z_{(n)}$ of an i.i.d. sample $(Z_1, \ldots, Z_n)$ of standard Gaussian variables. More formally, assume that $n \geq 2$ and consider the subset $\mathcal{K}_n$ of $\{1, 2, \ldots, n/2\}$ defined by

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq [\log_2(n/2)]\}. \tag{5}$$

The test statistics is then defined as

$$\Psi_{2,\alpha} = \sup_{k \in \mathcal{K}_n} \mathbb{1}_{\{Z^{(v_n)}_{(n-k+1)} > q_{\alpha_n,k}\}}, \tag{6}$$

where for all $u \in]0,1[$, $q_{u,k}$ denotes the $(1-u)$ quantile of $Z_{(n-k+1)}$ and

$$\alpha_n = \sup\left\{u \in]0,1[,\ \mathbb{P}_0\left(\exists k \in \mathcal{K}_n, Z_{(n-k+1)} < q_{u,k}\right) \leq \alpha\right\}. \tag{7}$$

In some sense, we proceed to a multiple testing approach. The subset $\mathcal{K}_n$ is related to the different orders that are included in the detection process: we do not use the whole ordered sample in order to enhance the performances of our test. The term α_n then corresponds to a correction of the level of each individual test $\mathbb{1}_{\{Z^{(v_n)}_{(n-k+1)} > q_{\alpha,k}\}}$ that guarantees a final first kind error of level α. The quantiles $q_{\alpha,k}$ can be explicitly computed. We refer to Laurent et al. (2014) for more details regarding the construction of this testing procedure.

Theorem 5 enhances the performances of the test $\Psi_{2,\alpha}$.

Theorem 5. *Let $\alpha, \beta \in]0,1[$ be fixed. Then, the testing procedure $\Psi_{2,\alpha}$ introduced in (6) is of level α. Moreover, there exists a positive constant $C(\alpha, \beta, M)$ depending only on α, β and M such that*

$$\sup_{\substack{f \in \mathcal{F}_2[M] \\ \varepsilon\|\mu\| \geq \rho}} \mathbb{P}_f\left(\Psi_{2,\alpha} = 0\right) \leq \beta,$$

for all $\rho \in \mathbb{R}_+^*$ such that

$$\rho \geq \rho^\dagger := C(\alpha, \beta, M) \frac{d^{1/4}}{\sqrt{n}} \sqrt{\ln \ln n}$$

provided that

$$n\varepsilon \geq C(\alpha, \beta, M).$$

The main conclusion of the above result is that the test $\Psi_{2,\alpha}$ based on the ordered statistics has a maximal second kind error bounded by β as soon as $\rho \geq \rho^{\dagger}$. Hence, $\Psi_{2,\alpha}$ exhibits essentially the same level of performances than $\Psi_{1,\alpha}$. The only difference with the bound displayed in Theorem 4 is related to an additional log-term ($\sqrt{\ln \ln n}$). Indeed, we proceed to a multiple testing procedure since we consider several indices for the ordered statistics. This log-term corresponds to the price to pay for such a construction. However, we stress that it could be removed provided that the upper bound M on $\|\mu\|$ is known.

This testing procedure has been included in the present study since we expect some robustness properties w.r.t. futur extensions of this work (see Section 4 for an extended discussion).

3.3. A testing procedure for the alternative class $\mathcal{F}_{\infty}[M]$

In the previous section, we have applied the procedure proposed in Laurent et al. (2014) on the projection of the data on a given axis (generated by the empirical mean of the sample). Here, we alternatively consider a projection on the canonical axis. In other word, we apply the procedure of Laurent et al. (2014) on each canonical direction in order to detect a possible contamination. As summarized in Theorem 6 below, this approach appears to be convenient when dealing with the alternative class $\mathcal{F}_{\infty}[M]$.

Let $(e_1, \ldots, e_d)$ be the canonical basis of $\mathbb{R}^d$. In the following, we set

$$Z_{ij} := \langle X_i, e_j \rangle \quad \forall i \in \{1, \ldots, n\},\ j \in \{1, \ldots d\}.$$

Then, for a given $j \in \{1, \ldots, d\}$, we can remark that $(Z_{i,j})_{i=1\ldots n}$ is a unidimensional sample and we denote by $(Z_{(i),j})_{i=1\ldots n}$ its associated ordered statistics. Then, for a given level $\alpha \in]0,1[$, we set

$$T^{+}_{\alpha,j} := \sup_{k \in \mathcal{K}_n} \left\{ \mathbb{1}_{Z_{(n-k+1),j} > q_{\alpha_n,k}} \right\},$$

and

$$T^{-}_{\alpha,j} := \sup_{k \in \mathcal{K}_n} \left\{ \mathbb{1}_{Z_{(k),j} < -q_{\alpha_n,k}} \right\},$$

where $\mathcal{K}_n$, α_n and $q_{\alpha,k}$ are defined as in Equations (5) and (7). Then, we consider the following test statistics

$$\Psi_{3,\alpha} = \sup_{1 \leq j \leq d} \max \left(T^{+}_{\frac{\alpha}{2d},j}, T^{-}_{\frac{\alpha}{2d},j} \right). \tag{8}$$

The following theorem provides a control on the first and second kind errors of the test $\Psi_{3,\alpha}$ when the alternative is measured via the infinite norm.

Theorem 6. *Let $\alpha, \beta \in]0,1[$ be fixed. Then, the test $\Psi_{3,\alpha}$ introduced in (8) is of level α. Moreover, if $n \geq 3$ and*

$$8.25 \times \frac{\ln\left[4d\log_2(n/2)/\alpha\right]}{n} \leq \int_M^{+\infty} \phi_1(x)dx,$$

then, there exists a positive constant $C(\beta, M)$ depending only on β and M, such that

$$\sup_{\substack{f \in \mathcal{F}_\infty[M] \\ \varepsilon\|\mu\|_\infty \geq \rho}} \mathbb{P}_f(\Psi_{3,\alpha} = 0) \leq \beta,$$

as soon as

$$\rho \geq C(\beta, M) \times \sqrt{\frac{\ln\ln(n)\ln(d/\alpha)}{n}}. \tag{9}$$

Theorem 6, together with Theorem 3 allow to characterize the separation frontier for the testing problem (1) when the energy (norm of μ) is measured w.r.t. the infinite norm. As discussed in Section 2, the problem appears to be easier in this setting as the dimension of the data grows: the price to pay is a term of order $\sqrt{\ln(d)}$ (against $d^{1/4}$ with the l_2-norm).

By the way, the construction of the test $\Psi_{3,\alpha}$ highlights the presence of this log-term: we proceed to $2d$ different tests (for each dimension, in the directions e_j and $-e_j$), and reject H_0 as soon as one of these tests detects something. This exactly corresponds to a multiple testing approach. The price to pay relies in the Bonferroni correction in each test: α is replaced by $\alpha/2d$, which implies the presence of a $\sqrt{\ln(d)}$ term in the separation condition.

4. Discussion

In our opinion, the main contribution of this paper is a sharp characterization of the role played by the dimension in a two-component mixture detection context. As discussed above, the price to pay in a multidimensional setting is a term of order $d^{1/4}$ (resp. $\sqrt{\ln(d)}$) when the energy in the alternative is measured w.r.t. the l_2-norm (resp. l_∞-norm). At this step, several questions are still open and provide possible outcomes for futur investigations.

First of all, according to the classical denomination in the statistical literature, our investigations have been drawn in a *dense* regime. Indeed, the proportion parameter ε is not allowed to be (asymptotically) smaller than $1/\sqrt{n}$. On the other hand, when

$d = 1$, several analyses have been conducted in a so-called *sparse* regime, i.e. when $\varepsilon \ll 1/\sqrt{n}$. In this context, it could be challenging to investigate the testing problem (1) in a *sparse* context and to precisely determine the influence of the dimension d on the problem. By the way, it seems necessary to propose a procedure that will be convenient for any considered regime, i.e. in some sense adaptive to the asymptotic of parameter ε.

Several additional investigations could be driven in this setting, among them: considering more general benchmark distributions (i.e. different from the standard Gaussian distribution), heteroscedastic mixtures or taking into account some uncertainty on the reference distribution. All these questions are outside the scope of the paper but could be at the core of future contributions.

5. Proof of Theorems 2 and 3

For the sake of convenience, we introduce the subset $\mathcal{F}[\rho, M]$ which corresponds to

$$\mathcal{F}_2[\rho, M] = \{f \in \mathcal{F}_2[M]; \varepsilon\|\mu\| \geq \rho\}$$

in the first proof, and

$$\mathcal{F}_\infty[\rho, M] = \{f \in \mathcal{F}_\infty[M]; \varepsilon\|\mu\|_\infty \geq \rho\}$$

in the second proof, for any given radius $\rho > 0$. Following Ingster and Suslina (2003) or Baraud (2002), we will use a Bayesian argument in order to bound the minimax separation radius in the two contexts. Thus, we consider a subset $\{g_\omega; \omega \in \Omega\}$ of $\mathcal{F}[\rho, M]$ which will be specify for each proof later. Then,

$$\sup_{f \in \mathcal{F}[\rho,M]} \mathbb{P}_f(\Psi_\alpha = 0) \geq \mathbb{P}_{g_\omega}(\Psi_\alpha = 0), \forall \omega \in \Omega.$$

Denoting the uniform probability measure π on the finite set Ω, we have

$$\sup_{f \in \mathcal{F}[\rho,M]} \mathbb{P}_f(\Psi_\alpha = 0) \quad \geq \quad \int_\Omega \mathbb{P}_{g_\omega}(\Psi_\alpha = 0) d\pi(\omega) := \mathbb{P}_\pi(\Psi_\alpha = 0). \tag{10}$$

Using (10) and similar computations as in Ingster and Suslina (2003) or Baraud (2002), we obtain

$$\begin{aligned} \inf_{\Psi_\alpha} \sup_{f \in \mathcal{F}[\rho,M]} \mathbb{P}_f(\Psi_\alpha = 0) &\geq \inf_{\Psi_\alpha} \mathbb{P}_\pi(\Psi_\alpha = 0) \\ &\geq 1 - \alpha - \frac{1}{2}\sqrt{\mathbb{E}_0[L_\pi^2(\underline{X})] - 1}, \end{aligned}$$

where $L_\pi(\underline{X}) = \frac{d\mathbb{P}_\pi}{d\mathbb{P}_0}(\underline{X})$ is the likelihood ratio. In particular, if we can ensure that

$$\mathbb{E}_0\left[L_\pi^2(\underline{X})\right] < 1 + \eta(\alpha,\beta)^2,$$

where $\eta(\alpha,\beta) = 2(1-\alpha-\beta)$ for all $\alpha,\beta \in]0,1[$, then

$$\inf_{\Psi_\alpha} \sup_{f\in\mathcal{F}[\rho,M]} \mathbb{P}_f(\Psi_\alpha = 0) > 1 - \alpha - \frac{1}{2}\eta(\alpha,\beta) = \beta.$$

In the two following proofs displayed below, we will specify the subset $\{g_\omega; \omega \in \Omega\}$ and propose an upper bound for the term $\mathbb{E}_0[L_\pi^2(\underline{X})]$.

5.1. Proof of Theorem 2

In this proof, recall that

$$\mathcal{F}[\rho,M] = \mathcal{F}_2[\rho,M] = \{f \in \mathcal{F}_2[M]; \varepsilon\|\mu\| \geq \rho\},$$

for any given radius $\rho > 0$.

We now consider $r \in]0,M]$ and $\varepsilon \in]0,1[$ such that $\varepsilon r = \rho$. In this context, we choose $\Omega = \{-1,1\}^d$ and

$$\forall \omega \in \Omega,\ g_\omega(.) = (1-\varepsilon)\phi_d(.) + \varepsilon\phi_d\left(. - \frac{r}{\sqrt{d}}\omega\right) \in \mathcal{F}_2[\rho,M].$$

Then, we have to propose an upper bound for the term $\mathbb{E}_0[L_\pi^2(\underline{X})]$ where in this setting

$$\begin{aligned} L_\pi(\underline{X}) &= \frac{d\mathbb{P}_\pi}{d\mathbb{P}_0}(\underline{X}) \\ &= \frac{1}{2^d}\sum_{\omega\in\{-1,1\}^d}\prod_{i=1}^n\left[(1-\varepsilon) + \varepsilon\frac{\phi_d(X_i - \frac{r}{\sqrt{d}}\omega)}{\phi_d(X_i)}\right] \\ &= \frac{1}{2^d}\sum_{\omega\in\{-1,1\}^d}\prod_{i=1}^n\left[(1-\varepsilon) + \varepsilon e^{-\frac{r^2}{2}}e^{\langle X_i, \frac{r}{\sqrt{d}}\omega\rangle}\right]. \end{aligned}$$

Thus,

$$\begin{aligned} L_\pi^2(\underline{X}) = \frac{1}{2^{2d}}\sum_{\omega,\tilde{\omega}\in\{-1,1\}^d}\prod_{i=1}^n\Big[&(1-\varepsilon)^2 + \varepsilon(1-\varepsilon)e^{-\frac{r^2}{2}}\left(e^{\langle X_i, \frac{r}{\sqrt{d}}\omega\rangle} + e^{\langle X_i, \frac{r}{\sqrt{d}}\tilde{\omega}\rangle}\right) \\ &+\varepsilon^2 e^{-r^2}e^{\langle X_i, \frac{r}{\sqrt{d}}(\omega+\tilde{\omega})\rangle}\Big]. \end{aligned}$$

Since for all $\mu \in \mathbb{R}^d$, $\mathbb{E}_0\left[e^{\langle X_i,\mu\rangle}\right] = e^{\|\mu\|^2/2}$, we have

$$\mathbb{E}_0[L_\pi^2(\underline{X})] = \frac{1}{2^{2d}} \sum_{\omega,\tilde{\omega}\in\{-1,1\}^d} \prod_{i=1}^n \left[(1-\varepsilon)^2 + 2\varepsilon(1-\varepsilon) + \varepsilon^2 e^{-r^2} e^{\frac{r^2}{2d}\|\omega+\tilde{\omega}\|^2}\right].$$

Noticing that $\|\omega+\tilde{\omega}\|^2 = 2d + 2\langle\omega,\tilde{\omega}\rangle$,

$$\begin{aligned}
\mathbb{E}_0[L_\pi^2(\underline{X})] &= \frac{1}{2^{2d}} \sum_{\omega,\tilde{\omega}\in\{-1,1\}^d} \prod_{i=1}^n \left[1 - \varepsilon^2 + \varepsilon^2 e^{\frac{r^2}{d}\langle\omega,\tilde{\omega}\rangle}\right] \\
&= \frac{1}{2^{2d}} \sum_{\omega,\tilde{\omega}\in\{-1,1\}^d} \left[1 + \varepsilon^2\left(e^{\frac{r^2}{d}\langle\omega,\tilde{\omega}\rangle} - 1\right)\right]^n \\
&= \mathbb{E}\left[\left\{1 + \varepsilon^2\left(e^{\frac{r^2}{d}\langle W,\tilde{W}\rangle} - 1\right)\right\}^n\right],
\end{aligned}$$

where W and $\tilde{W}$ are two independent d-dimensional Rademacher random variables, i.e.

$$\mathbb{P}(W = w) = \mathbb{P}(\tilde{W} = w) = \frac{1}{2^d} \quad \forall w \in \{-1,1\}^d.$$

Noticing that

$$\langle W, \tilde{W}\rangle = \sum_{j=1}^d W_j \tilde{W}_j$$

and that the variables $W_j\tilde{W}_j$ for $1 \le j \le d$ are also i.i.d. Rademacher random variables, $\langle W, \tilde{W}\rangle$ has the same distribution as $Y = \sum_{j=1}^d W_j$. This leads to

$$\mathbb{E}_0[L_\pi^2(\underline{X})] = \mathbb{E}\left[\left\{1 + \varepsilon^2\left(e^{\frac{r^2}{d}Y} - 1\right)\right\}^n\right].$$

We now use the following inequality which holds for any real number u such that $|u| \le M$:

$$|e^u - 1 - u| \le \frac{e^M}{2} u^2. \tag{11}$$

Since $\left|r^2\frac{Y}{d}\right| \le r^2 \le M^2$, we have

$$e^{\frac{r^2}{d}Y} - 1 \le \frac{r^2}{d}Y + \frac{e^{M^2}}{2}\frac{r^4}{d^2}Y^2.$$

Hence, we have

$$0 \leq 1-\varepsilon^2 \leq 1+\varepsilon^2\left(e^{\frac{r^2}{d}Y}-1\right) \leq 1+\frac{\varepsilon^2 r^2}{\sqrt{d}}\left(\frac{Y}{\sqrt{d}}+\frac{e^{M^2}}{2}M^2\frac{Y^2}{d}\right).$$

Setting $C(M)=1+e^{M^2}M^2/2$,

$$0 \leq 1+\varepsilon^2\left(e^{\frac{r^2}{d}Y}-1\right) \leq 1+C(M)\frac{\varepsilon^2 r^2}{\sqrt{d}}\left(\frac{|Y|}{\sqrt{d}}\vee\frac{Y^2}{d}\right),$$

which leads to

$$\mathbb{E}_0[L_\pi^2(\underline{X})] \leq \mathbb{E}\left[\left\{1+a\left(\frac{|Y|}{\sqrt{d}}\vee\frac{Y^2}{d}\right)\right\}^n\right],$$

where

$$a = C(M)\varepsilon^2 r^2/\sqrt{d}. \tag{12}$$

Using the inequality $\ln(1+x)\leq x$ for all $x\geq 0$, we have

$$\begin{aligned}\mathbb{E}_0[L_\pi^2(\underline{X})] &\leq \mathbb{E}\left[e^{\left\{na\left(\frac{|Y|}{\sqrt{d}}\vee\frac{Y^2}{d}\right)\right\}}\right],\\ &\leq e^{na}\mathbb{P}\left(\frac{|Y|}{\sqrt{d}}\leq 1\right)+\mathbb{E}\left[e^{na\frac{Y^2}{d}}\mathbb{1}_{\left\{\frac{|Y|}{\sqrt{d}}>1\right\}}\right].\end{aligned}$$

Moreover, using an integration by part

$$\mathbb{E}\left[e^{na\frac{Y^2}{d}}\mathbb{1}_{\left\{\frac{|Y|}{\sqrt{d}}>1\right\}}\right] \leq e^{na}\mathbb{P}\left(\frac{|Y|}{\sqrt{d}}>1\right)+\int_{e^{na}}^{+\infty}\mathbb{P}\left(e^{na\frac{Y^2}{d}}>t\right)dt,$$

leading to

$$\mathbb{E}_0[L_\pi^2(\underline{X})] \leq e^{na}+\int_{e^{na}}^{+\infty}\mathbb{P}\left(e^{na\frac{Y^2}{d}}>t\right)dt.$$

We deduce from Hoeffding's inequality that for all $x>0$,

$$\mathbb{P}\left(\frac{|Y|}{\sqrt{d}}>x\right)\leq 2\exp\left(-x^2/2\right).$$

Hence, for all $t>e^{na}$,

$$\mathbb{P}\left(e^{na\frac{Y^2}{d}}>t\right)\leq 2t^{-1/2na}.$$

In the particular case where $na < 1/2$, we get

$$\begin{aligned}\mathbb{E}_0[L_\pi^2(\underline{X})] &\leq e^{na} + 2\int_{e^{na}}^{+\infty} t^{-1/2na} dt, \\ &\leq e^{na}\left(1 + \frac{4na}{1-2na}e^{-1/2}\right) \leq h(na),\end{aligned}$$

where the function $h(.)$ is defined as

$$h(x) = e^x\left(1 + \frac{4x}{1-2x}e^{-1/2}\right) \quad \forall x \in [0, 1/2[.$$

The function h is non decreasing on $[0, 1/2[$. Hence

$$na \leq 1/4 \Rightarrow \mathbb{E}_0[L_\pi^2(\underline{X})] \leq h(1/4) \leq 3 < 1 + \eta(\alpha, \beta)^2,$$

since, according to our assumption

$$\alpha + \beta < 1 - \frac{1}{\sqrt{2}} \simeq 0.293 \Rightarrow (1 - \alpha - \beta)^2 > 1/2.$$

In order to conclude the proof, just remark from (12) that

$$na \leq 1/4 \Leftrightarrow \varepsilon^2 r^2 \leq \sqrt{d}/(4C(M)n).$$

Hence, setting $(\rho^\#)^2 = \sqrt{d}/(4C(M)n)$, we get that if $\rho < \rho^\#$, then $\mathbb{E}_0[L_\pi^2(\underline{X})] < 1 + \eta(\alpha, \beta)^2$, which leads to the desired result.

□

5.2. Proof of Theorem 3

In this context,

$$\mathcal{F}[\rho, M] = \mathcal{F}_\infty[\rho, M] := \{f \in \mathcal{F}_\infty[M]; \varepsilon\|\mu\|_\infty \geq \rho\},$$

for any $\rho > 0$. Let $r \in]0, M]$ such that $\varepsilon r = \rho$. In this context, we choose

$$\Omega = \left\{\omega \in \{0,1\}^d \ s.t. \ \sum_{j=1}^d \omega_j = 1\right\},$$

and we define for all $\omega \in \Omega$,

$$g_\omega(.) = (1-\varepsilon)\phi_d(.) + \varepsilon\phi_d(. - r\omega) \in \mathcal{F}_\infty[\rho, M].$$

Now, we turn our attention to the control of the associated likelihood ratio. For each $j = 1, \ldots, d$, let $D^{(j)} \in \{0,1\}^d$ such that $D_\ell^{(j)} = \mathbb{1}_{\ell=j}$ and

$$\begin{aligned} L_\pi(\underline{X}) &= \frac{d\mathbb{P}_\pi}{d\mathbb{P}_0}(\underline{X}), \\ &= \left[\prod_{i=1}^n \phi_d(X_i)\right]^{-1} \left[\frac{1}{d}\sum_{j=1}^d \prod_{i=1}^n \left\{(1-\varepsilon)\phi_d(X_i) + \varepsilon\phi_d(X_i - rD^{(j)})\right\}\right], \\ &= \frac{1}{d}\sum_{j=1}^d U_j(\underline{X}), \end{aligned}$$

with

$$U_j(\underline{X}) = \prod_{i=1}^n \left\{(1-\varepsilon) + \varepsilon\frac{\phi_d(X_i - rD^{(j)})}{\phi_d(X_i)}\right\}.$$

Thus

$$\mathbb{E}_0[L_\pi^2(\underline{X})] = \frac{1}{d^2}\sum_{j=1}^d \mathbb{E}_0[U_j(\underline{X})^2] + \frac{1}{d^2}\sum_{k\neq j} \mathbb{E}_0[U_j(\underline{X})U_k(\underline{X})]. \tag{13}$$

In a first time, we can remark that for all $j \in \{1, \ldots, d\}$

$$\begin{aligned} \mathbb{E}_0[U_j(\underline{X})^2] &= \mathbb{E}_0\left[\left(\prod_{i=1}^n \left\{(1-\varepsilon) + \varepsilon\frac{\phi_d(X_i - rD^{(j)})}{\phi_d(X_i)}\right\}\right)^2\right], \\ &= \mathbb{E}_{\phi_d}\left[\left\{(1-\varepsilon) + \varepsilon\frac{\phi_d(X_1 - rD^{(j)})}{\phi_d(X_1)}\right\}^2\right]^n, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}_{\phi_d}\left[\left\{(1-\varepsilon) + \varepsilon\frac{\phi_d(X - rD^{(j)})}{\phi_d(X)}\right\}^2\right] \\ &= (1-\varepsilon)^2 + \varepsilon^2\int_{\mathbb{R}^d}\frac{\phi_d^2(x - rD^{(j)})}{\phi_d(x)}dx + 2(1-\varepsilon)\varepsilon\int_{\mathbb{R}^d}\phi_d(x - rD^{(j)})dx, \\ &= (1-\varepsilon)^2 + \varepsilon^2 e^{r^2} + 2(1-\varepsilon)\varepsilon, \\ &= 1 + \varepsilon^2(e^{r^2} - 1), \end{aligned}$$

since $\int_{\mathbb{R}^d} \frac{\phi_d^2(x-\mu)}{\phi_d(x)}dx = \exp(\|\mu\|^2)$. Thus

$$\mathbb{E}_0[U_j(\underline{X})^2] = \left\{1+\varepsilon^2(e^{r^2}-1)\right\}^n.$$

Concerning the second sum in (13), we obtain for all $j,k \in \{1,\ldots,d\}$, $j \neq k$

$$\begin{aligned}
&\mathbb{E}_0[U_j(\underline{X})U_k(\underline{X})] \\
&= \mathbb{E}_0\left[\prod_{i=1}^n \left\{(1-\varepsilon)+\varepsilon\frac{\phi_d(X_i-rD^{(j)})}{\phi_d(X_i)}\right\}\left\{(1-\varepsilon)+\varepsilon\frac{\phi_d(X_i-rD^{(k)})}{\phi_d(X_i)}\right\}\right], \\
&= \left\{\mathbb{E}_{\phi_d}\left[(1-\varepsilon)^2+(1-\varepsilon)\varepsilon\frac{\phi_d(X_1-rD^{(j)})+\phi_d(X_1-rD^{(k)})}{\phi_d(X_1)}\right.\right. , \\
&\qquad\qquad \left.\left. +\varepsilon^2\frac{\phi_d(X_1-rD^{(j)})\phi_d(X_1-rD^{(k)})}{\phi_d(X_1)^2}\right]\right\}^n, \\
&= \left\{(1-\varepsilon)^2+2(1-\varepsilon)\varepsilon+\varepsilon^2\exp[r^2\langle D^{(j)},D^{(k)}\rangle]\right\}^n, \\
&= \left\{(1-\varepsilon)^2+2(1-\varepsilon)\varepsilon+\varepsilon^2\right\}^n = 1,
\end{aligned}$$

since $\int_{\mathbb{R}^d} \frac{\phi_d(x-\mu_1)\phi_d(x-\mu_2)}{\phi_d(x)}dx = \exp(\langle\mu_1,\mu_2\rangle)$. Finally,

$$\mathbb{E}_0[L_\pi^2(\underline{X})] = \frac{1}{d}\left\{1+\varepsilon^2(e^{r^2}-1)\right\}^n + \frac{d(d-1)}{d^2}.$$

We obtain

$$\begin{aligned}
\mathbb{E}_0[L_\pi^2(\underline{X})] < 1+\eta(\alpha,\beta)^2 \quad &\Leftrightarrow \quad \frac{1}{d}\left\{1+\varepsilon^2(e^{r^2}-1)\right\}^n + \frac{d(d-1)}{d^2} < 1+\eta(\alpha,\beta)^2, \\
&\Leftrightarrow \quad \left\{1+\varepsilon^2(e^{r^2}-1)\right\}^n < 1+d\eta(\alpha,\beta)^2, \\
&\Leftrightarrow \quad \varepsilon^2(e^{r^2}-1) < \exp\left[\frac{1}{n}\ln(1+d\eta(\alpha,\beta)^2)\right]-1.
\end{aligned}$$

Since $0 < r \leq M$,

$$\varepsilon^2(e^{r^2}-1) \leq C(M)(\varepsilon r)^2 = C(M)(\rho^\star)^2,$$

where the constant $C(M)$ satisfies $C(M) = (1+M^2e^{M^2}/2)$ (see (11)). On the other hand,

$$\exp\left[\frac{1}{n}\ln(1+d\eta(\alpha,\beta)^2)\right]-1 > \frac{1}{n}\ln(1+d\eta(\alpha,\beta)^2).$$

Thus, the condition $\mathbb{E}_0[L_\pi^2(\underline{X})] < 1+\eta(\alpha,\beta)^2$ is fulfilled as soon as

$$C(M)\rho^2 < \frac{1}{n}\ln(1+d\eta(\alpha,\beta)^2)$$

which is equivalent to

$$\rho < \sqrt{\frac{1}{C(M)}\frac{1}{n}\ln(1+d\eta(\alpha,\beta)^2)}.$$

This concludes the proof of Theorem 3.

□

6. Proof of Theorems 4, 5 and 6

6.1. Proof of Theorem 4

First, remark that

$$\|\sqrt{n}\bar{X}_n\|^2 = \sum_{j=1}^{d}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_{ij}\right)^2.$$

Under H_0, X_{ij} are i.i.d. standard Gaussian random variables. Hence $\|\sqrt{n}\bar{X}_n\|^2$ is a chi-squared random variable with d degrees of freedom and

$$\mathbb{P}_0(\Psi_{1,\alpha}=1) = \mathbb{P}_0(\|\sqrt{n}\bar{X}_n\|^2 > \upsilon_\alpha) = \alpha,$$

according to the definition of the quantile υ_α. The test $\Psi_{1,\alpha}$ is hence of level α.

Now, we want to control the second kind error. Under H_1, each variable X_i can be written as

$$X_i = V_i\mu + \eta_i,$$

where V_i is a Bernoulli variable with parameter ε, $\eta_i \sim \mathcal{N}_d(0_d, I_d)$ and V_i and η_i are independent. Then

$$\sqrt{n}\bar{X}_n = \frac{S}{\sqrt{n}}\mu + B$$

where $S = \sum_{i=1}^{n} V_i \sim \mathcal{B}(n,\varepsilon)$ is a binomial random variable with parameters (n,ε), $B = \sum_{i=1}^{n}\eta_i/\sqrt{n} \sim \mathcal{N}_d(0_d, I_d)$ and S, B are independent. In particular, conditionally to S, the variable $\|\sqrt{n}\bar{X}_n\|^2 = \left\|\frac{S}{\sqrt{n}}\mu + B\right\|^2$ has a non-central chi-squared distribution with d degrees of freedom and noncentrality parameter $\lambda_S = \|\frac{S}{\sqrt{n}}\mu\|^2$. Introduce

$$h_S = d + \lambda_S - 2\sqrt{[d+2\lambda_S]\ln(2/\beta)}.$$

According to Lemma 2 in Appendix A (see also Laurent et al., 2012)

$$\mathbb{P}\left(\left\|\frac{S}{\sqrt{n}}\mu + B\right\|^2 \leq h_S \;\middle|\; S\right) \leq \frac{\beta}{2}.$$

Hence, for each $f \in \mathcal{F}_2[M]$,

$$\begin{aligned}
\mathbb{P}_f\left(\Psi_{1,\alpha}=0\right) &= \mathbb{P}_f\left(\|\sqrt{n}\bar{X}_n\|^2 \leq \upsilon_\alpha\right), \\
&= \mathbb{P}\left(\left\|\frac{S}{\sqrt{n}}\mu + B\right\|^2 \leq \upsilon_\alpha\right), \\
&= \mathbb{P}\left(\left\{\left\|\frac{S}{\sqrt{n}}\mu + B\right\|^2 \leq \upsilon_\alpha\right\} \cap \{h_S \leq \upsilon_\alpha\}\right) \\
&\qquad + \mathbb{P}\left(\left\{\left\|\frac{S}{\sqrt{n}}\mu + B\right\|^2 \leq \upsilon_\alpha\right\} \cap \{h_S > \upsilon_\alpha\}\right), \\
&\leq \mathbb{P}\left(h_S \leq \upsilon_\alpha\right) + \mathbb{P}\left(\left\|\frac{S}{\sqrt{n}}\mu + B\right\|^2 \leq h_S\right), \\
&\leq \mathbb{P}\left(h_S \leq \upsilon_\alpha\right) + \frac{\beta}{2}.
\end{aligned}$$

According to Lemma 1 in Appendix A,

$$\upsilon_\alpha \leq d + b(\alpha, d) \quad \text{where} \quad b(\alpha, d) = 2\ln(1/\alpha) + 2\sqrt{d\ln(1/\alpha)}.$$

Hence

$$\begin{aligned}
\mathbb{P}(h_S \leq \upsilon_\alpha) &\leq \mathbb{P}(h_S \leq d + b(d,\alpha)), \\
&\leq \mathbb{P}(\lambda_S - 2\sqrt{[d+2\lambda_S]\ln(2/\beta)} \leq b(d,\alpha)), \\
&\leq \mathbb{P}(\lambda_S - 2\sqrt{2\ln(2/\beta)}\sqrt{\lambda_S} - [2\sqrt{d\ln(2/\beta)} + b(d,\alpha)] \leq 0), \\
&\leq \mathbb{P}(\sqrt{\lambda_S} \leq R(\alpha,\beta,d)),
\end{aligned}$$

with

$$R(\alpha,\beta,d) = \sqrt{2\ln(2/\beta)} + \sqrt{2\ln(2/\beta) + 2\sqrt{d\ln(2/\beta)} + b(\alpha,d)}.$$

We notice that $R(\alpha,\beta,d) \leq C(\alpha,\beta)d^{1/4}$ where $C(\alpha,\beta)$ is a constant depending only on α and β. Assuming that $\sqrt{n}\varepsilon\|\mu\| > C(\alpha,\beta)d^{1/4}$ and using a Tchebychev's

inequality leads to

$$\begin{aligned}\mathbb{P}(h_S \leq \upsilon_\alpha) &\leq \mathbb{P}\left(S \leq \frac{\sqrt{n}}{\|\mu\|}C(\alpha,\beta)d^{1/4}\right),\\ &\leq \mathbb{P}\left(|S-n\varepsilon| > n\varepsilon - \frac{\sqrt{n}}{\|\mu\|}C(\alpha,\beta)d^{1/4}\right),\\ &\leq \frac{n\varepsilon\|\mu\|^2}{[n\varepsilon\|\mu\| - \sqrt{n}C(\alpha,\beta)d^{1/4}]^2},\\ &\leq \frac{n\varepsilon\|\mu\|M}{[n\varepsilon\|\mu\| - \sqrt{n}C(\alpha,\beta)d^{1/4}]^2}.\end{aligned}$$

Then, $\mathbb{P}(h_S \leq \upsilon_\alpha) \leq \beta/2$ is satisfied if

$$\frac{n\varepsilon\|\mu\|}{[n\varepsilon\|\mu\| - \sqrt{n}C(\alpha,\beta)d^{1/4}]^2} \leq \frac{\beta}{2M}.$$

The last inequality is fulfilled if

$$\varepsilon\|\mu\| \geq \frac{\sqrt{n}C(\alpha,\beta)d^{1/4}}{n} + \frac{M}{\beta n} + \frac{\sqrt{1+2\sqrt{n}C(\alpha,\beta)d^{1/4}\beta/M}}{\beta n/M}$$

noticing that $\frac{nx}{(nx-c)^2} \leq B$ is fulfilled if and only if $x \notin \left[\frac{c}{n} + \frac{1}{2Bn} \pm \frac{\sqrt{1+4cB}}{2Bn}\right]$.
Thus $\mathbb{P}(h_S \leq \upsilon_\alpha) \leq \beta/2$ if

$$\varepsilon\|\mu\| \geq \frac{\tilde{C}(\alpha,\beta,M)d^{1/4}}{\sqrt{n}},$$

where $\tilde{C}(\alpha,\beta,M)$ is a positive constant depending on α,β and M.

□

6.2. Proof of Theorem 5

Following the definition of α_n, $\Psi_{2,\alpha}$ is ensured to be a level-α test. Then, under H_1, each variable Y_i can be written as

$$Y_i = V_i\mu + \eta_i, \quad i = 1\dots n,$$

where $V_i \sim \mathcal{B}(\varepsilon)$ denotes a random Bernoulli variable and $\eta_i \sim \mathcal{N}_d(0_d, I_d)$, V_i and η_i being independent. Hence, for all $i \in \{1,\dots,n\}$, we get

$$Z_i^{(\upsilon_n)} = \langle Y_i, \upsilon_n\rangle = V_i\langle\mu,\upsilon_n\rangle + \langle\eta_i,\upsilon_n\rangle$$

which satisfies

$$Z_i^{(v_n)}|v_n \sim (1-\varepsilon)\mathcal{N}(0,1) + \varepsilon\mathcal{N}(\langle \mu, v_n\rangle, 1).$$

Let $f \in \mathcal{F}_2[M]$ and Ω_{v_n} be the event defined as

$$\Omega_{v_n} = \left\{ \varepsilon\langle \mu, v_n\rangle > C\left(\frac{\beta}{2}, M\right)\frac{\kappa_n}{\sqrt{n}}\sqrt{\ln(1/\alpha)} \right\},$$

where $\kappa_n = \sqrt{\ln\ln(n)}$ and $C(\beta, M)$ denotes the constant appearing in (20) (see Appendix B). In the following, the constant $C(\frac{\beta}{2}, M)\sqrt{\ln(1/\alpha)}$ is denoted $C_{\alpha,\beta,M}$.

Then, using Lemma 3 in Appendix B, we get

$$\begin{aligned}
\mathbb{P}_f(\Psi_{2,\alpha}=0) &\leq \mathbb{E}_{v_n}\left[\mathbb{E}_f\left[\mathbb{1}_{\Psi_{2,\alpha}=0}\mathbb{1}_{\Omega_{v_n}}|v_n\right]\right] + \mathbb{P}_f(\Omega_{v_n}^c) \\
&\leq \mathbb{E}_{v_n}\left[\mathbb{1}_{\Omega_{v_n}}\mathbb{E}_f\left[\mathbb{1}_{\Psi_{2,\alpha}=0}|v_n\right]\right] + \mathbb{P}_f(\Omega_{v_n}^c) \\
&\leq \frac{\beta}{2} + \mathbb{P}_f(\Omega_{v_n}^c). \qquad (14)
\end{aligned}$$

Now, we turn our attention to the control of $\mathbb{P}_f(\Omega_{v_n}^c)$. Using the definition of v_n we obtain

$$\begin{aligned}
\mathbb{P}_f(\Omega_{v_n}^c) &= \mathbb{P}_f\left(\varepsilon\langle \mu, v_n\rangle \leq \frac{C_{\alpha,\beta,M}}{\sqrt{n}}\kappa_n\right), \\
&= \mathbb{P}_f\left(\varepsilon\left\langle \mu, \frac{\bar{A}_n}{\|\bar{A}_n\|}\right\rangle \leq \frac{C_{\alpha,\beta,M}}{\sqrt{n}}\kappa_n\right), \\
&= \mathbb{P}_f\left(\varepsilon\left\langle \mu, \bar{A}_n\right\rangle \leq \frac{C_{\alpha,\beta,M}}{n}\|\sqrt{n}\bar{A}_n\|\kappa_n\right).
\end{aligned}$$

At this step, remark that the variable $\bar{A}_n$ can be written as

$$\bar{A}_n = \frac{S}{n}\mu + \frac{U}{\sqrt{n}}, \qquad (15)$$

where $S \sim \mathcal{B}(n,\varepsilon)$, $U \sim \mathcal{N}_d(0_d, I_d)$, and S and U are independent. Moreover, conditionally to S, $\|\sqrt{n}\bar{A}_n\|^2$ has a noncentral chi-squared distribution with d degrees of freedom and noncentrality parameter $\lambda_S = \|\frac{S}{\sqrt{n}}\mu\|^2$. Introduce

$$h(S) = d + \lambda_S + 2\sqrt{(d+2\lambda_S)x_\beta} + 2x_\beta$$

with $x_\beta = \ln(4/\beta)$. According to Lemma 2,

$$\mathbb{P}_f\left(\|\sqrt{n}\bar{A}_n\|^2 > h(S) \mid S\right) \leq \frac{\beta}{4}.$$

Then,

$$\begin{aligned}
\mathbb{P}_f(\Omega_{v_n}^c) &= \mathbb{P}_f\left(\varepsilon\left\langle\mu,\bar{A}_n\right\rangle \leq \frac{C_{\alpha,\beta,M}}{n}\|\sqrt{n}\bar{A}_n\|\kappa_n\right), \\
&\leq \mathbb{P}_f\left(\varepsilon\left\langle\mu,\bar{A}_n\right\rangle \leq \frac{C_{\alpha,\beta,M}}{n}\sqrt{h(S)}\kappa_n\right) + \mathbb{P}_f\left(\|\sqrt{n}\bar{A}_n\|^2 > h(S)\right), \\
&\leq \mathbb{P}_f\left(\varepsilon\left\langle\mu,\bar{A}_n\right\rangle \leq \frac{C_{\alpha,\beta,M}}{n}\sqrt{h(S)}\kappa_n\right) + \frac{\beta}{4} \\
&\leq \mathbb{P}_f\left(\frac{\varepsilon\|\mu\|^2}{n}S + \frac{\varepsilon\|\mu\|}{\sqrt{n}}Z \leq \frac{C_{\alpha,\beta,M}}{n}\sqrt{h(S)}\kappa_n\right) + \frac{\beta}{4}
\end{aligned}$$

since

$$\varepsilon\langle\mu,\bar{A}_n\rangle = \frac{\varepsilon\|\mu\|^2}{n}S + \frac{\varepsilon\|\mu\|}{\sqrt{n}}Z, \tag{16}$$

where $Z \sim \mathcal{N}(0,1)$ and is independent of S.

Let $C_{1,\beta} = \sqrt{8/\beta}$. According to the Tchebychev's inequality,

$$\mathbb{P}\left(|S - n\varepsilon| > C_{1,\beta}\sqrt{n\varepsilon}\right) \leq \frac{\beta}{8}.$$

Thus,

$$\begin{aligned}
\mathbb{P}_f(\Omega_{v_n}^c) \leq \mathbb{P}_f\left(\left\{\frac{\varepsilon\|\mu\|^2}{n}S + \frac{\varepsilon\|\mu\|}{\sqrt{n}}Z \leq \frac{C_{\alpha,\beta,M}}{n}\sqrt{h(S)}\kappa_n\right\} \cap \left\{|S - n\varepsilon| \leq C_{1,\beta}\sqrt{n\varepsilon}\right\}\right) \\
+\frac{\beta}{8} + \frac{\beta}{4}.
\end{aligned}$$

Note that, if $|S - n\varepsilon| \leq C_{1,\beta}\sqrt{n\varepsilon}$ and $\sqrt{n\varepsilon} > 2C_{1,\beta}$ then $S \in \left[\frac{1}{2}n\varepsilon, \frac{3}{2}n\varepsilon\right]$. Hence, on the event $|S - n\varepsilon| \leq C_{1,\beta}\sqrt{n\varepsilon}$

$$\begin{aligned}
\sqrt{h(S)} &= \left(d + \lambda_S + 2\sqrt{(d+2\lambda_S)x_\beta} + 2x_\beta\right)^{\frac{1}{2}} \\
&\leq \sqrt{d + 2x_\beta + 2\sqrt{dx_\beta}} + \frac{S\|\mu\|}{\sqrt{n}} + \sqrt{\frac{2\sqrt{2x_\beta}S\|\mu\|}{\sqrt{n}}} \\
&\leq a_{1,\beta}\sqrt{d} + \frac{3}{2}\sqrt{n}\varepsilon\|\mu\| + \sqrt{3\sqrt{2nx_\beta}\varepsilon\|\mu\|} \\
&\leq a_{1,\beta}\sqrt{d} + a_{2,\beta}\sqrt{n}\varepsilon\|\mu\|
\end{aligned}$$

since $S \leq \frac{3}{2}n\varepsilon$ and $\sqrt{n}\varepsilon\|\mu\| \geq 1$, where $a_{1,\beta}$ and $a_{2,\beta}$ are two positive constants. Then, as soon as $\sqrt{n\varepsilon} > 2C_{1,\beta}$,

$$\begin{aligned}
&\mathbb{P}_f(\Omega_{v_n}^c) \\
&\leq \mathbb{P}_f\left(\left\{\frac{\varepsilon\|\mu\|^2}{n}S + \frac{\varepsilon\|\mu\|}{\sqrt{n}}Z \leq \frac{C_{\alpha,\beta,M}}{n}\kappa_n\sqrt{h(S)}\right\} \cap \left\{|S - n\varepsilon| \leq C_{1,\beta}\sqrt{n\varepsilon}\right\}\right) + \frac{3\beta}{8}, \\
&\leq \mathbb{P}_f\left(\frac{1}{2}\varepsilon^2\|\mu\|^2 + \frac{\varepsilon\|\mu\|}{\sqrt{n}}Z \leq \frac{C_{\alpha,\beta,M}}{n}\kappa_n\left[a_{1,\beta}\sqrt{d} + a_{2,\beta}\sqrt{n}\varepsilon\|\mu\|\right]\right) + \frac{3\beta}{8}, \\
&\leq \mathbb{P}_f\left(\frac{\varepsilon\|\mu\|}{\sqrt{n}}Z \leq \frac{C_{\alpha,\beta,M}}{n}\kappa_n\left[a_{1,\beta}\sqrt{d} + a_{2,\beta}\sqrt{n}\varepsilon\|\mu\|\right] - \frac{1}{2}\varepsilon^2\|\mu\|^2\right) + \frac{3\beta}{8}.
\end{aligned}$$

Let $z_{\beta/8}$ be the $\beta/8$ quantile of the standard Gaussian distribution. Then,

$$\mathbb{P}\left(\frac{\varepsilon\|\mu\|}{\sqrt{n}}Z \leq \frac{\varepsilon\|\mu\|}{\sqrt{n}}z_{\beta/8}\right) = \frac{\beta}{8}.$$

Then, $\mathbb{P}_f(\Omega_{v_n}^c) \leq \frac{\beta}{8} + \frac{3\beta}{8} = \frac{\beta}{2}$ if

$$\begin{aligned}
&\frac{C_{\alpha,\beta,M}}{n}\kappa_n\left[a_{1,\beta}\sqrt{d} + a_{2,\beta}\sqrt{n}\varepsilon\|\mu\|\right] - \frac{1}{2}\varepsilon^2\|\mu\|^2 \leq \frac{\varepsilon\|\mu\|}{\sqrt{n}}z_{\beta/8} \\
\Leftrightarrow\ & n\varepsilon^2\|\mu\|^2 - 2C_{\alpha,\beta,M}\kappa_n\left[a_{1,\beta}\sqrt{d} + a_{2,\beta}\sqrt{n}\varepsilon\|\mu\|\right] + 2z_{\beta/8}\sqrt{n}\varepsilon\|\mu\| \geq 0 \\
\Leftrightarrow\ & n\varepsilon^2\|\mu\|^2 - 2v_{1,n}\sqrt{n}\varepsilon\|\mu\| - v_{2,n}\sqrt{d} \geq 0, \qquad (17)
\end{aligned}$$

with $v_{1,n} = C_{\alpha,\beta,M}\kappa_n a_{2,\beta} - z_{\beta/8} > 0$ and $v_{2,n} = 2C_{\alpha,\beta,M}\kappa_n a_{1,\beta} > 0$. Inequality (17) is fulfilled if

$$\sqrt{n}\varepsilon\|\mu\| \geq v_{1,n} + \sqrt{v_{1,n}^2 + v_{2,n}\sqrt{d}}.$$

Finally, $\mathbb{P}_f\left(\Psi_{2,\alpha} = 0\right) \leq \beta$ occurs as soon as

$$\sqrt{n}\varepsilon\|\mu\| > C^* d^{1/4}\kappa_n \geq 1 \text{ and } n\varepsilon \geq \tilde{C} \qquad (18)$$

for some constants C^* and $\tilde{C}$ which only depend on α, β and M.

□

6.3. Proof of Theorem 6

The test $\Psi_{3,\alpha}$ is a level-α test since

$$\begin{aligned}\mathbb{P}_0(\Psi_{3,\alpha}=1) &= \mathbb{P}_0\left(\exists j\in\{1,\dots,d\}; T^+_{\frac{\alpha}{2d},j}=1 \cup T^-_{\frac{\alpha}{2d},j}=1\right),\\ &\leq \sum_{j=1}^{d}\left[\mathbb{P}_0\left(T^+_{\frac{\alpha}{2d},j}=1\right)+\mathbb{P}_0\left(T^-_{\frac{\alpha}{2d},j}=1\right)\right],\\ &\leq d\times 2\times\frac{\alpha}{2d}=\alpha.\end{aligned}$$

We now consider the control on the second kind error of this test $\Psi_{3,\alpha}$ when the alternative is measured via the infinite norm. If

$$\varepsilon\|\mu\|_\infty \geq C(\beta,M)\times\sqrt{\frac{\ln\ln(n)\ln(2d/\alpha)}{n}},$$

then there exists $j_0\in\{1,\dots,d\}$ such that

$$\mathbb{P}_f\left(T^+_{\frac{\alpha}{2d},j_0}=0\right)\leq\beta \text{ or } \mathbb{P}_f\left(T^-_{\frac{\alpha}{2d},j_0}=0\right)\leq\beta$$

according to Lemma 3 in Appendix B. Then,

$$\begin{aligned}\mathbb{P}_f(\Psi_{3,\alpha}=0) &= \mathbb{P}_f\left(\sup_{1\leq j\leq d}\max\left(T^+_{\frac{\alpha}{2d},j},T^-_{\frac{\alpha}{2d},j}\right)=0\right),\\ &= \mathbb{P}_f\left(\bigcap_{j=1..d}\{T^+_{\frac{\alpha}{2d},j}=0\}\cap\{T^-_{\frac{\alpha}{2d},j}=0\}\right),\\ &\leq \inf_{j=1\dots d}\left[\mathbb{P}_f(T^+_{\frac{\alpha}{2d},j}=0)\wedge\mathbb{P}_f(T^-_{\frac{\alpha}{2d},j}=0)\right],\\ &\leq \mathbb{P}_f(T^+_{\frac{\alpha}{2d},j_0}=0)\wedge\mathbb{P}_f(T^-_{\frac{\alpha}{2d},j_0}=0)\leq\beta.\end{aligned}$$

□

Appendix A: Properties for chi-squared distribution and noncentral chi-squared distribution.

In this section, we present some well-known results there are useful throughout the proofs. The first lemma is concerned with deviation of a chi-squared random variable, proposed in Laurent and Massart (2000).

Lemma 1. *Let U be a chi-squared random variable with d degrees of freedom. Then,*

- *for any positive x,*

$$\begin{cases} \mathbb{P}(U \geq d + 2\sqrt{dx} + 2x) \leq e^{-x}, \\ \mathbb{P}(U \leq d - 2\sqrt{dx}) \leq e^{-x}. \end{cases}$$

- *For any given $\alpha \in]0,1[$, let $u(d,\alpha)$ be the $(1-\alpha)$-quantile of $\chi^2(d)$. Then*

$$u(d,\alpha) \leq d + 2\ln(1/\alpha) + 2\sqrt{d\ln(1/\alpha)} = d + b(\alpha, d).$$

This second lemma provides the control of deviations of a noncentral chi-squared random variable, available in Birgé (2001).

Lemma 2. *Let T be a noncentral chi-squared random variable with d degrees of freedom and a noncentrality parameter λ. Then, for any positive x,*

$$\begin{cases} \mathbb{P}\left(T \geq d + \lambda + 2\sqrt{(d+2\lambda)x} + 2x\right) \leq e^{-x}, \\ \mathbb{P}\left(T \leq d + \lambda - 2\sqrt{(d+2\lambda)x}\right) \leq e^{-x}. \end{cases}$$

Appendix B: Unidimensional test

We have previously proposed some testing procedures that uses results proposed in Laurent et al. (2014) in a unidimensional context. For the sake of convenience, we reproduce a slightly different version of theses contributions in order to facilitate the understanding of the proofs.

Let $(Z_1, \ldots, Z_n)$ be i.i.d. random variables from an unknown density g w.r.t. the Lebesgue measure on $\mathbb{R}$. The order statistics are denoted by $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)}$. We want to test

$$H_0 : \ g = \phi(.) \quad \text{versus} \quad H_1 : \ g \in \mathcal{G}_1[M],$$

where $\phi(.) = \phi_1(.)$ is the unidimensional Gaussian density and

$$\mathcal{G}_1[M] = \{z \in \mathbb{R} \mapsto (1-\varepsilon)\phi(z) + \varepsilon\phi(z-\tau); 0 < \tau \leq M\}.$$

Let $\alpha \in]0,1[$. Let T_α be the test statistics defined as

$$T_\alpha := \sup_{k \in \mathcal{K}_n} \left\{ \mathbb{1}_{Z_{(n-k+1)} > z_{\alpha_n,k}} \right\}, \tag{19}$$

where, for all $u \in]0,1[$, $z_{u,k}$ is the $(1-u)$-quantile of $Z_{(n-k+1)}$ under the null hypothesis and

$$\alpha_n = \sup \left\{ u \in]0,1[, \mathbb{P}_{H_0} \left(\exists k \in \mathcal{K}_n, Z_{(n-k+1)} > z_{u,k} \right) \leq \alpha \right\}.$$

The following lemma establishes sufficient conditions that allow to control the second kind error of T_α.

Lemma 3. *Let $\beta \in]0, 1-\alpha[$. Assume that $n \geq 3$ and*

$$8.25 \times \frac{\ln(2\log_2(n/2)/\alpha)}{n} \leq \int_M^{+\infty} \phi(x)dx.$$

Then, there exists a positive constant $C(\beta, M)$ depending only on β and M, such that if

$$\rho \geq C(\beta, M)\sqrt{\frac{\ln\ln(n)\ln(1/\alpha)}{n}}, \tag{20}$$

then,

$$\sup_{\substack{g \in \mathcal{G}_1[M] \\ \varepsilon\tau \geq \rho}} \mathbb{P}_g(T_\alpha = 0) \leq \beta.$$

Proof. By definition, the test statistics T_α introduced in (19) is exactly of level α, namely

$$\mathbb{P}_{H_0}(T_\alpha = 1) = \mathbb{P}_{H_0} \left(\exists k \in \mathcal{K}_n, Z_{(n-k+1)} > z_{\alpha_n,k} \right) \leq \alpha,$$

thanks to the definition of α_n.

In order to control the second kind error of the test T_α, we first give an upper bound for $z_{\alpha_n,k}$. Let $\bar{\Phi}(x) = 1-\Phi(x)$, where Φ is the cumulative distribution function associated to the density function ϕ. For all $\alpha \in]0,1[$ and $k \in \{1, 2, \ldots, n/2\}$, let $t_{\alpha,k}$ be the positive real number defined as

$$\bar{\Phi}\left(t_{\alpha,k}\right) = \frac{k}{n}\left[1 - \sqrt{\frac{2\ln(\frac{2}{\alpha})}{k}}\right] \tag{21}$$

if $k > 2\ln(\frac{2}{\alpha})$, and $t_{\alpha,k} = +\infty$ otherwise. According to Lemma A.1 in Laurent et al. (2014), $z_{\alpha_n,k} \leq t_{\alpha_n,k}$. Considering $g \in \mathcal{G}_1[M]$, we want to control the second kind

error of the test:

$$\begin{aligned}\mathbb{P}_g(T_\alpha = 0) &= \mathbb{P}_g\left(\forall k \in \mathcal{K}_n, Z_{(n-k+1)} \leq z_{\alpha_n,k}\right) \\ &\leq \inf_{k\in\mathcal{K}_n} \mathbb{P}_g\left(Z_{(n-k+1)} \leq z_{\alpha_n,k}\right). \end{aligned} \tag{22}$$

Since $z_{\alpha_n,k} \leq t_{\alpha_n,k}$, using Markov's inequality

$$\begin{aligned}\mathbb{P}_g\left(Z_{(n-k+1)} \leq z_{\alpha_n,k}\right) &\leq \mathbb{P}_g(Z_{(n-k+1)} \leq t_{\alpha_n,k}) \\ &\leq \mathbb{P}_g\left(\sum_{i=1}^{n}\left\{\mathbb{1}_{\{Z_i\leq t_{\alpha_n,k}\}} - q_1\right\} > n(1-q_1) - k\right) \\ &\leq \frac{n(1-q_1)}{[n(1-q_1)-k]^2}\end{aligned}$$

if $n(1-q_1) - k > 0$, where

$$1 - q_1 = \mathbb{P}_g\left(Z_1 \geq t_{\alpha_n,k}\right) = (1-\varepsilon)\bar{\Phi}\left(t_{\alpha_n,k}\right) + \varepsilon\bar{\Phi}\left(t_{\alpha_n,k} - \tau\right).$$

Note that the inequality $\frac{nx}{(nx-k)^2} \leq \beta$ is fulfilled if and only if $x \notin \left[\frac{k}{n} + \frac{2}{\beta n} \pm \frac{2\sqrt{1+4k\beta}}{\beta n}\right]$. Then,

$$\mathbb{P}_f\left(Z_{(n-k+1)} < t_{\alpha_n,k}\right) \leq \beta,$$

if

$$(1-\varepsilon)\bar{\Phi}\left(t_{\alpha_n,k}\right) + \varepsilon\bar{\Phi}\left(t_{\alpha_n,k} - \tau\right) > \frac{k}{n} + \frac{2}{n\beta} + \frac{2\sqrt{1+4k\beta}}{n\beta}. \tag{23}$$

Now, we consider $k \in \mathcal{K}_n$ such that

$$\frac{0.99}{2}\bar{\Phi}(M) \leq \frac{k}{n} \leq 0.99\ \bar{\Phi}(M).$$

The set of solutions of this inequation is not empty since under the assumptions of Lemma 3, $0.99\ \bar{\Phi}(M)n \geq 1$. Note that $|\mathcal{K}_n| \leq \log_2(n/2)$, hence $\alpha_n \geq \alpha/|\mathcal{K}_n| \geq \alpha/\log_2(n/2)$. We will show that Condition (23) is fulfilled. Using a Taylor expansion at the order 1,

$$(1-\varepsilon)\bar{\Phi}\left(t_{\alpha_n,k}\right) + \varepsilon\bar{\Phi}\left(t_{\alpha_n,k} - \tau\right) = \bar{\Phi}\left(t_{\alpha_n,k}\right) + \varepsilon\tau\phi(a)$$

where a belongs to the interval $]t_{\alpha_n,k} - \tau, t_{\alpha_n,k}[$. We recall that $\bar{\Phi}\left(t_{\alpha_n,k}\right) = \frac{k}{n}\left[1 - \sqrt{\frac{2\ln(2/\alpha_n)}{k}}\right]$. Using (21), we just have to show that

$$\varepsilon\tau\phi(a) \geq \frac{2}{\beta n} + \frac{2\sqrt{1+4k\beta}}{\beta n} + \frac{\sqrt{k}}{n}\sqrt{2\ln(2/\alpha_n)}, \tag{24}$$

in order to prove that Condition (23) holds.

Next, we want to prove that $[t_{\alpha_n,k} - \tau, t_{\alpha_n,k}]$ remains included in a fixed interval $[c_1(M), c_2(M)]$ with $c_1(M) > 0$.
On one hand, we have

$$t_{\alpha_n,k} \geq \bar{\Phi}^{-1}\left(\frac{k}{n}\right) \geq \bar{\Phi}^{-1}\left(0.99\ \bar{\Phi}(M)\right),$$

and

$$t_{\alpha_n,k} - M \geq \bar{\Phi}^{-1}\left(0.99\ \bar{\Phi}(M)\right) - M := c_1(M) > 0.$$

Moreover,

$$\begin{aligned}\bar{\Phi}\left(t_{\alpha_n,k}\right) &\geq \frac{0.99}{2}\bar{\Phi}(M) - \sqrt{\frac{2\ln(2/\alpha_n)}{n}}\sqrt{0.99\ \bar{\Phi}(M)} \\ &\geq \frac{\bar{\Phi}(M)}{200}\end{aligned}$$

since $(8.25)\ln(2\log_2(n/2)/\alpha)/n \leq \bar{\Phi}(M)$. This implies that

$$t_{\alpha_n,k} \leq \bar{\Phi}^{-1}\left(\bar{\Phi}(M)/200\right) := c_2(M).$$

Finally, since $\phi(a) \geq \phi(c_2(M))$, (24) is satisfied if $\varepsilon\tau \geq C(\beta, M)\sqrt{\ln\ln(n)\ln(1/\alpha)/n}$ for some suitable constant $C(\alpha, \beta, M)$.

□

Acknowledgements

This work was supported by the French Agence Nationale de la Recherche (ANR-13-JS01-0001-01, project MixStatSeq).

References

Arias-Castro, E., Candes, E., and Durand, A. (2011). Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39:278–304.

Azaïs, J.-M., Gassiat, É., and Mercadier, C. (2009). The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM Probab. Stat.*, 13:301–327.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

Birgé, L. (2001). *An alternative point of view on Lepski's method*, volume Volume 36 of *Lecture Notes–Monograph Series*, pages 113–133. Institute of Mathematical Statistics.

Butucea, C. and Ingster, Y. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652Ñ2688.

Cai, T. T., Jeng, X. J., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(5):629–662.

Cai, T. T., Jin, J., and Low, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449.

Cai, T. T. and Wu, Y. (2014). Optimal detection for sparse mixtures against a given null distribution. To appear in IEEE Transactions on Information Theory.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J. Statist. Plann. Inference*, 43(1-2):19–40.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994.

Garel, B. (2007). Recent asymptotic results in testing for mixtures. *Comput. Statist. Data Anal.*, 51(11):5295–5304.

Ingster, Y. (1999). Minimax detection of a signal for l^n-balls. *Mathematical Methods of Statistics*, 7(4):401–428.

Ingster, Y. and Suslina, I. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Laurent, B., Loubes, J.-M., and Marteau, C. (2012). Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electronic Journal of Statistics*, 6:91–122.

Laurent, B., Marteau, C., and Maugis-Rabusseau, C. (2014). Non-asymptotic detection of mixtures with unknown mean. *To appear in Bernoulli*.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley series in Probability and Statistics.

Verzelen, N. and Arias-Castro, E. (2014). Detection and feature selection in sparse mixture models. *ArXiv 1405.1478*.